基于混合互信息算法的文本情感分析 *

王　义，戴月明

(江南大学 物联网工程学院, 江苏 无锡 214122)

摘　要：针对互信息（mutual information, MI）特征选择方法存在的正负相关性的现象以及未考虑特征项在不同类别内词频的问题，提出了一种混合互信息特征选择算法（hybrid mutual information, HMI）。该算法引入逆文档频率系数和类间词频信息系数，使得整个文档中的词频信息以及每个类之间的词频信息得以有效利用；引入正负相关性系数，区分正相关性和负相关性，并进行有效的利用。通过实验对比表明，混合互信息算法可以有效地提高特征选择的质量，进而提高文本情感分析的效果。

关键词：互信息；特征选择；正负相关性；词频信息；情感分析

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2018.08.0537

Text sentiment analysis based on hybrid mutual information algorithm

Wang Yi, Dai Yueming

(*School of Internet of Things Engineering*, *Jiangnan University*, *Wuxi* Jiangsu 214122, *China*)

Abstract: Aiming at the phenomenon of positive and negative correlation in the feature selection method of mutual information (MI) and the problem of not considering the word frequency of the feature items in different categories, a hybrid mutual information feature selection algorithm (HMI) is proposed. By introducing the inverse document frequency coefficient and the inter-class word frequency information coefficient, the algorithm can effectively utilize the word frequency information in the whole document and the word frequency information between each class. The positive and negative correlation coefficient is introduced to distinguish positive correlation and negative correlation and to make effective use. The experimental results show that the hybrid mutual information algorithm can effectively improve the quality of feature selection and then improve the effect of text emotional analysis.

Key words: mutual information; feature selection; positive and negative correlation; word frequency information; sentiment classification

0　引言

随着科技的不断发展，互联网的普及越来越高，对于数据分析的需求也日益增长，一些关于商品以及服务的评论越来越多。因此，从这些评论以及评价中对其观点进行情感分析成为了当下热门研究方向[1]。文本情感分析是分析文本的情感倾向，并在文本中挖掘作者的观点、态度等有效信息[2]，因此文本情感分析也被称为文本观点挖掘。目前，基于情感词典以及基于机器学习是文本情感分析的两种主要方法，而基于机器学习的方法是当前情感分类的主流方法[3]。在文本情感分析中，文本数据一般被表示为空间向量模型（VSM）[4]，借助此模型，可以将文本数据转换成为结构化数据，以便计算机能够对其进行处理。在一般的数据集中，特征项通常会达到上万个特征，稍大的数据集，甚至会达到上百万个特征，因此如何在降低特征空间的维度的同时，提高文本情感分类的效果，就成为文本情感分析中的关键问题[5]。特征选择自然也就成为文本情感分析中的一个重要部分。

特征选择的主要目的就是为了通过去除噪声[6]，选择出高质量，具有代表性的词汇，从而提高分类的准确率。目前，常见的特征选择方法包括：卡方统计量（Chi-square statistic，CHI）、文档频数（document frequency，DF）、信息增益（information gain，IG）、互信息（mutual information，MI）等。在这些特征选择算法中，互信息以其时间复杂度低,易于理解以及使用便捷等优点，成为了一种重要的特征选择方法[7]。但是，传统互信息算法因为其没考虑词频因素，导致选取的特征词的质量较低[8]；另外，在计算特征项的互信息值时，忽略了呈负相关性的特征项，导致负相关性的特征值会明显削弱总体的特征值，从而降低了互信息算法的精确性。

综上所述，针对互信息算法在特征选择时的不足，本文引入了逆文档频率系数,类间频率系数以及正负相关性系数，提出一种混合互信息（hybrid mutual information，HMI）特征选择方法。通过理论分析以及实验证明，该算法能有效地利用词频信息以及正负相关性信息提高特征选择的质量，从而提高情感分类的精确度。

1　互信息特征选择方法

互信息（MI）是一种基于统计学的算法，一般用来度量两个统计量之间的相互关联程度[9]。而在文本情感分析中，互信息算法一般被用来计算文本中的特征项与各个类别之间的关联程度。当特征项与该类别的关联程度越大时，该特征项与该类的互信息值就越大，则说明该特征项对于该类就越具有代表性 [9]。记 t_k 为特征项的集合，$k=1,2,\ldots,m$；c_j 为训练集类别的集合，$j=1,2,\ldots,r$；则 t_k 与 c_j 之间的互信息值的计算公式如下：

收稿日期：2018-08-02；**修回日期：**2018-09-28　**基金项目：**国家自然科学基金资助项目（61572237）

作者简介：王义（1994-），男，安徽芜湖人，硕士研究生，主要研究方向为自然语言处理、机器学习（425498661@qq.com）；戴月明（1964-），男，江苏常熟人，副教授，硕士，主要研究方向为人工智能、软件工程.

$$MI(t_k,c_j)=\log\frac{p(t_k,c_j)}{p(t_k)p(c_j)}=\log\frac{p(t_k\mid c_j)}{p(t_k)} \tag{1}$$

其中:$p(t_k,c_j)$表示训练集中既包含特征 t_k 又属于类别的文本 c_j 概率，$p(t_k)$表示在整个文本训练集中包含特征 t_k 的文本概率，$p(c_j)$表示文本属于训练集中类别 c_j 的文本概率，$p(t_k|c_j)$表示类别 c_j 中，包含特征 t_k 的文本概率。对于多个类别的训练集，特征项 t_k 与训练集中各个类别的互信息计算公式如下所示：

$$MI(t_k)=\sum_{j=1}^{r}p(c_j)\log\frac{p(t_k,c_j)}{p(t_k)p(c_j)}=\sum_{j=1}^{r}p(c_j)\log\frac{p(t_k\mid c_j)}{p(t_k)} \tag{2}$$

当特征选择时，$MI(t_k)$的互信息值按照从大到小排序，根据维度的选取需求，选取其中较大的前 n 项值所对应的特征项进行文本的向量的表示。

互信息算法计算的是包含特征项 t_k 的文本数量与训练集中每个类别 c_j 中文本数量的之比，最重要的特征就是考虑了不同特征项和这一类别的同现频率，有效地利用了文本的类别信息[10]。但是互信息方法也存在着一些明显的不足之处，例如在式（2）中，各个特征项在不同类别之间的频数的差异并没有体现出来，也没有考虑包含特征项文本训练集频数之间的联系。在文本情感分析中，特征项与各个类别之间的相关性分为正相关性和负相关性，正相关的特征项在文本情感分类中起主要的作用[11]，但负相关特征对于最终的文本情感分类结果也有着重要的作用。从式（2）中可以看出，正负相关性的互信息值相互抵消了，从而忽略了负相关性的作用。

2　混合互信息特征选择方法

2.1　类间词频信息系数

通过以上分析，在传统的互信息特征选择方法中，只考虑了类内特征项的频率[12]，但是特征项的词频数在文本情感分析中的作用不仅仅体现在类内，它也在类和类之间起着非常重要的作用。如果一个特征项对于某一类的代表能力越强，那么该特征项应当集中在某一类中，也就是在这一类中该特征项的词频应当较大，相反在其他类中应当尽量少的出现。假定特征项 t_k 为类别 c_j 的特征项，那么在特征项 t_k 在类别 c_j 中应尽可能多的出现，而在其他类别 $c_q(q\neq j)$中应尽可能的少出现。那么在理论推导中，对于类别代表能力较强的特征项，在不同类别之间的标准差应当尽可能大。基于以上考虑，本文在式（2）的基础之上，引入类间词频信息系数 α，则 α 的定义形式所示为

$$\alpha=\sqrt{\frac{1}{m}\sum_{j=1}^{m}\left[tf_j(t_k)-\frac{1}{m}\sum_{i=1}^{m}tf_i(t_k)\right]^2} \tag{3}$$

其中:$tf_j(t_k)$表示为特征 t_k 在类别 j 中出现的频数，m 表示为类别的总数。那么，式（2）则可以写成

$$MI(t_k)=\alpha\times\sum_{j=1}^{r}p(c_j)\log\frac{p(t_k\mid c_j)}{p(t_k)} \tag{4}$$

式（4）中，在互信息方法中引入类间词频信息系数 α，系数 α 统计了特征项在每一类间的词频的标准差，使得特征项的词频信息在不同类别中得以体现，提高了互信息特征选择方法的效率。因此，该式进一步提高了文本情感分类的效果。

2.2　逆文档频率系数

上文通过引入类间词频信息系数 α，了解到当一个特征项集中出现在某一类的文本中时，则它对文本类别的代表能力就越强。在传统的互信息方法中，还忽略了出现特征项的文档频率，例如"他""你""是"这样的词在很多文本中都有可α能会出现，然而这类词对于文本的区别能力并不高[13]。因此，如果一个特征项出现在大多数的文本中，那么意味着这个特征项对于文本类别的区分能力就越弱。基于以上所述，为了对此类特征项进行区分，增加特征项的区分力，引入逆文档系数 β，来调节这类问题。β 的定义形式如下所示：

$$\beta=\ln\frac{N}{f(t_k)+0.01} \tag{5}$$

那么，可以将式（4）写成

$$MI(t_k)=\alpha\times\sum_{j=1}^{r}\beta\times p(c_j)\log\frac{p(t_k\mid c_j)}{p(t_k)} \tag{6}$$

其中，N 代表训练集中的文档总数，$f(t_k)$代表包含特征项 t_k 的文本数量，分母后加上小数 0.01 确保分母不为 0，以保证系数的有效性。在式（5）中，由于 $N\geq f(t_k)$恒成立，当存在更多包含特征项 t_k 的文本时，也就是 $f(t_k)$越大时，$f(t_k)$越接近 N，系数 β 越接近 0，那么，逆文档系数 β 值对于该特征项的 MI 值的影响就越小。通过引入逆文档频率系数，降低了一些常用词作为特征项对于最后分类结果的影响，提高了特征选择的效率。

2.3　正负相关性系数

由式（1）可以看出，在计算特征项 t_k 和类别 c_j 的互信息值时，当 $p(t_k|c_j)$大于 $p(t_k)$，此时 $MI(t_k)>0$，这表明特征项 t_k 和类别 c_j 是正相关的。说明当 $p(t_k|c_j)$的值越大，而 $p(t_k)$的值越小时，特征项 t_k 所能代表类别 c_j 的能力就越强。而当 $p(t_k|c_j)$小于 $p(t_k)$，此时 $MI(t_k)<0$，这表明特征项 t_k 和类别 c_j 之间是负相关的。说明当 $p(t_k|c_j)$的值越小，而 $p(t_k)$的值越大时，特征项 t_k 与类别 c_j 之间的信息量就越少，特征项 t_k 代表类别 c_j 的能力就越低。从式（2）中可以看出，当计算类别集合的 MI 值时，特征项与类别为负相关性的部分值会削弱该特征项最终的 MI 值。在文本情感分类中，正相关性特征有利于提高最终的准确率，而负相关性特征有利于提高最终的查全率[14]，因此负相关性特征的作用也不能忽视[15]。针对这一现象，本文引入正负相关性系数 γ 来调节互信息方法中出现的正负相关性问题。

那么，在类别 c_j（j=1,2,…,r）中，首先定义：

$$\overline{f(t_k)}=\frac{1}{m}\sum_{j=1}^{m}f_j(t_k) \tag{7}$$

其中：$\overline{f(t_k)}$ 表示每个类别中含有的特征项 t_k 的平均文本数，$f_j(t_k)$代表类别 c_j 中包含特征项 t_k 的文本数量。

当 $p(t_k|c_j)$大于 $p(t_k)$时，γ 的定义形式如下所示：

$$\gamma=\omega\times\frac{f_j(t_k)-\overline{f(t_k)}}{\overline{f(t_k)}} \tag{8}$$

当 $p(t_k|c_j)$大于 $p(t_k)$时，γ 的定义形式如下所示：

$$\gamma=(1-\omega)\times\frac{f_j(t_k)-\overline{f(t_k)}}{\overline{f(t_k)}} \tag{9}$$

其中：ω 为调节因子，ω 的理论取值范围为 0.1~0.9，用来调节正负相关特征项的影响力，使得特征项无论是正相关还是负相关，都充分发挥其对于最终情感分类的作用。另外，在 γ 中，当特征项与类别呈现负相关时，$p(t_k|c_j)$的值越大，而 $p(t_k)$的值越小时，说明特征项在该类中出现的次数较少，因此，系数 γ 中的$(f_j(t_k)-\overline{f(t_k)})$很好地应对了这一情况，进而有

效利用了与类别项呈负相关性的特征项。$(f_j(t_k)-\overline{f(t_k)})/\overline{f(t_k)}$ 也相应表示了特征项文本集合的偏离程度，对于提高最终的特性选择效率有了一定的提高。

综上所述，混合互信息（HMI）的定义形式如下所示：

$$HMI(t_k)=\alpha\times\sum_{j=1}^{r}\beta\times\gamma\times p(c_j)\log\frac{p(t_k\mid c_j)}{p(t_k)} \tag{10}$$

HMI 算法的伪代码如下：

```
for each document dj ∈D do
    for each word tk ∈dj do
        IF word∈ Cj then
            tc++    //所求特征项在类别 Cj 中出现的频数
        end if
    end for
end for
for each category Cj ∈C do
    Ck++    //数据集类别总数
end for
for each document dj∈D do
    if word in Cj then
        dk++    //类别 Cj 中包含所求特征项的文本数量
    end if
end for
for each document dj∈Cj do
    Dj++    //类别 Cj 中文本的数量
end for
for each document dj∈D do
    if word∈dj then
        count++    //含所求特征项的文本数量
    end if
end for
for each document dj∈D do
    N++    //文本总数
end for
α=sqrt (square (tc-(sum (tk)/Ck))/Ck)
β=log [N/ (count+0.01)]
if (dk /Dj) >= (count/N) then
    γ=ω*[(dk -count/Ck)/ (count/Ck)]
else
    γ= (1- ω)*[(dk -count/Ck)/ (count/Ck)]
end if
HMI (tk ,Cj) =MI*α*β*γ
```

3　实验结果及分析

3.1　实验语料集

本文实验分别采用来自于谭松波的酒店管理评论语料集以及美的空调评论语料集来进行实验，两者语料集分别有 4 000 条评论数据，分为正向评论和负向评论两类 ，其中正向（pos）评论 2 000 条，负向（neg）评论 2000 条。为验证算法的有效性，文本采用交叉实验的方式，将语料库的数据取其中的 80%作为训练集，用于训练，其余 20%作为测试集，用于检验分类器的效果，进行实验分析。表 1 展示了本文所用语料集中训练集的相应实例，表 2 展示了本文所用语料集中测试集的相应实例。

表 1　训练集语料实例

Table 1 training set corpus examples

训练集语料实例：
S1：设备很简单服务意识很差餐厅的设施就跟街边的小滩一样，卫生条件很不好。
S2：房间很小，装修太差，电线全部裸露在外面，脏，不是一般的脏。
S3：位置不错，在市中心，周围吃饭等很方便，房间一如既往的干净。

表 2　测试集语料实例

Table 2 Test set corpus examples

测试集语料实例：
S4：地点和位置很好，晚上比较安静，设施较全。对于一般的自助游来说比较合适。
S5：酒店还可以，早餐也不错，值得推荐，若价格能再低点就更好了。
S6：房间装修陈旧，下水管堵塞，晚上折腾了 2 个多小时，还是没有修好。
S7：预定的房间给的是问题房，水泵声音太吵，地段离风景点偏了点，虫子很多。

3.2　实验评价标准

在文本情感分析中，比较常用的评价标准有查准率（precision），也叫做准确率，和查全率（recall），也叫做召回率，以及综合了准确率和查全率的评价标准 F_1 值。本文采用以上三种评价标准来对实验结果进行评价。文本情感分类的判断情况主要分为以下四种情况，如表 3 所示。

表 3　文本情感分类判断

Table 3 Judgment of text emotion classification

	正向样例	负向样例
预测结果为正向样例	TP	FP
预测结果为负向样例	FN	TN

表 4 列出了表 2 中测试集语料实例的实际结果和预测结果实验对比。

表 4　测试集语料实例实验对比

Table 4 Test set sample data

测试集语料实例	实际结果	预测结果
S4	正向	正向
S5	正向	负向
S6	负向	正向
S7	负向	负向

其中，TP 指的是预测为正向样例，实际也为正向样例的文本数，如表 4 中语料实例 **S4** 所示；FP 指的是预测为正向样例，实际为负向样例的文本数，如表 4 中语料实例 **S6** 所示；FN 指的是预测为负向样例，实际为正向样例的文本数，如表 4 中语料实例 **S5** 所示；TN 指的是预测为负向样例，实际也为负向样例的文本数，如表 4 中语料实例 **S7** 所示。

那么，关于准确率和召回率的定义形式为

$$Precision=\frac{TP}{TP+FP} \tag{11}$$

$$\mathrm{Re}call=\frac{TP}{TP+FN} \tag{12}$$

F_1 值则将准确率和召回率综合起来进行评价，其定义形式如下：

$$F1=\frac{2\times Precision\times \mathrm{Re}call}{Precision+\mathrm{Re}call} \tag{13}$$

3.3　实验步骤

a）对数据集进行预处理，对语料集进行标注，将正向语料和负向语料合并为一个文档，进行分词处理，本文采用的

是较为常用的 jieba 中文分词工具。针对表 1 训练集语料实例分词后结果如表 5 所示。

表 5 Jieba 分词处理后结果

Table 5 Results after Jieba participle processing

分词处理后结果：
S1: 设备 很 简单 服务 意识 很差 餐厅 的 设施 就 跟 街边 的 小滩 一样 ， 卫生条件 很 不好 。
S2: 房间 很小 ， 装修 太 差 ， 电线 全部 裸露 在 外面 ， 脏 ， 不是 一般 的 脏 。
S3: 位置 不错 ， 在 市中心 ， 周围 吃饭 等 很 方便 ， 房间 一如既往 的 干净 。

b）对文本中的停用词，标点符号等对文本情感分类无关的因素进去去除。针对表 1 训练集语料实例分词后结果如下表 6 所示。

表 6 去停用词后结果

Table 6 results after discontinuation of words

去停用词后结果：
S1: 设备 很 简单 服务 意识 很差 餐厅 设施 街边 小滩 卫生条件 很 不好
S2: 房间 很小 装修 太 差 电线 全部 裸露 外面 脏 不是 一般 脏
S3: 位置 不错 市中心 周围 吃饭 很 方便 房间 一如既往 干净

c）分别采用互信息（MI）、混合互信息（HMI）、卡方统计量（CHI）三种特征选择方法进行特征选择。

d）采用词袋模型（BOW）对特征项进行表示，并使用空间向量模型（VSM）将文本数据转换为结构化数据。

e）由于支持向量机（support vector machine ,SVM）分类器具有结构较为简单，全局最优等优点，已逐渐成为文本情感分析中的主流分类器。因此，本文实验采用了支持向量机分类器对数据进行训练以及测试，并对三种特征选择方法实验结果进行分析对比。

具体的实验流程如图 1 所示。

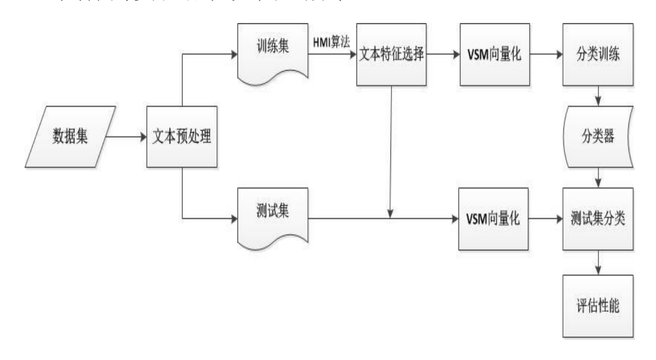


图 1 实验流程图

Fig. 1 Flow chart of experiment

3.4 结果及分析

本文实验中分别采用卡方统计量（CHI）、互信息（MI）以及本文提出的混合互信息（HMI）这三种不同的特征选择方法对数据集进行特征选择并进行对比实验。实验中，特征项表示方法采用 BOW 词袋模型，分别计算在 2000、3000、4000、5000、6000、7000、8000 和 9000 维度下的准确率（precision）、召回率（recall）以及 F_1 值。另外，由于正相关特征项起主要作用，对式（8）中的调节因子 ω 分别取 0.5、0.6、0.7、0.8 以及 0.9，通过多次对比实验，最终发现 ω 取 0.8 时，实验效果最佳。表 7~9 分别为酒店管理评论在不同维度下准确率、召回率以及 F_1 值的数据对比表格。

表 7 不同维度下准确率对比结果

Table 7 comparison of accuracy in different dimensions

特征维数	MI	HMI	CHI
2000	0.58	0.83	0.66
3000	0.61	0.83	0.71
4000	0.65	0.84	0.72
5000	0.70	0.83	0.74
6000	0.74	0.84	0.73
7000	0.79	0.85	0.74
8000	0.81	0.86	0.75
9000	0.81	0.86	0.75

表 8 不同维度下召回率对比结果

Table 8 comparison of recall rates in different dimensions

特征维数	MI	HMI	CHI
2000	0.59	0.90	0.78
3000	0.63	0.90	0.80
4000	0.68	0.89	0.77
5000	0.71	0.90	0.79
6000	0.77	0.88	0.79
7000	0.80	0.88	0.80
8000	0.81	0.89	0.80
9000	0.81	0.90	0.81

表 9 不同维度下 F1 值对比结果

Table 9 comparison of F1 values in different dimensions

特征维数	MI	HMI	CHI
2000	0.59	0.87	0.72
3000	0.63	0.87	0.75
4000	0.67	0.86	0.74
5000	0.70	0.86	0.76
6000	0.76	0.87	0.76
7000	0.80	0.86	0.77
8000	0.81	0.87	0.77
9000	0.81	0.87	0.78

由表 7~9 可以看出，混合互信息（HMI）算法在酒店管理评论数据集中，无论是在准确率、召回率，还是在 F_1 值上，这三个指标均明显优于另外两种特征选择方法。其中，由表 7 可以看出，准确率较 MI 算法提高了 5%，较 CHI 算法提高了 11%；由表 8 可以看出，召回率较 MI 算法提高了 9%，较 CHI 算法提高了 9%；由表 9 可以看出，F_1 值较 MI 算法提高了 6%，较 CHI 算法提高了 9%。可以看出，HMI 特征选择算法对于文本情感分类效果具有显著的提高。

表 10~12 分别为美的空调评论在不同维度下准确率、召回率以及 F1 值的数据对比表格。

表 10 不同维度下准确率对比结果

Table 10 comparison of accuracy in different dimensions

特征维数	MI	HMI	CHI
2000	0.55	0.69	0.56
3000	0.58	0.73	0.56
4000	0.64	0.75	0.61
5000	0.69	0.80	0.62
6000	0.71	0.83	0.64
7000	0.73	0.83	0.67
8000	0.71	0.84	0.71
9000	0.71	0.83	0.72

表 11　不同维度下召回率对比结果

Table 11 comparison of recall rates in different dimensions

特征维数	MI	HMI	CHI
2000	0.67	0.82	0.78
3000	0.69	0.85	0.80
4000	0.77	0.89	0.81
5000	0.79	0.90	0.83
6000	0.78	0.91	0.83
7000	0.78	0.91	0.85
8000	0.76	0.90	0.85
9000	0.78	0.90	0.84

表 12　不同维度下准确率对比结果

Table 12 comparison of accuracy in different dimensions

特征维数	MI	HMI	CHI
2000	0.60	0.79	0.69
3000	0.68	0.79	0.69
4000	0.73	0.80	0.70
5000	0.74	0.83	0.69
6000	0.74	0.80	0.70
7000	0.74	0.82	0.73
8000	0.73	0.82	0.72
9000	0.75	0.83	0.74

由表 10~12 可以看出，混合互信息（HMI）算法在美的空调评论中，无论是在准确率、召回率，还是在 F_1 值上，这三个指标均也同样明显优于另外两种特征选择方法。其中，由表 10 可以看出，准确率较 MI 算法提高了 12%，较 CHI 算法提高了 11%；由表 11 可以看出：召回率较 MI 算法提高了 12%，较 CHI 算法提高了 6%；由表 12 可以看出，F_1 值较 MI 算法提高了 8%，较 CHI 算法提高了 9%。综上可以看出，HMI 特征选择算法对于两种数据集的文本情感分类效果均具有显著的提高。

为了对不同特征维度下文本情感分类的准确率、召回率以及 F_1 值的变化情况有一个更加直观的认知，如图 2~4 所示，采用折线图来展现在酒店管理评论数据集中，不同维度下三个值的变化情况。

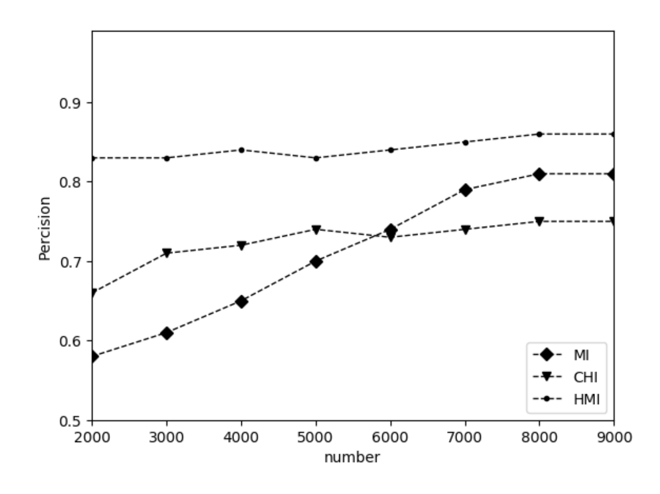


图 2　不同特征选择维度下准确率折线图

Fig. 2　Broken line diagram of accuracy under different feature selection dimensions

从图 2~4 可以看出，MI 和 CHI 算法效果在维度在 7000 时开始趋于稳定，而 HMI 算法自始至终都保持在一定的范围，且 HMI 算法整体的分类效果均优于另外两个算法。

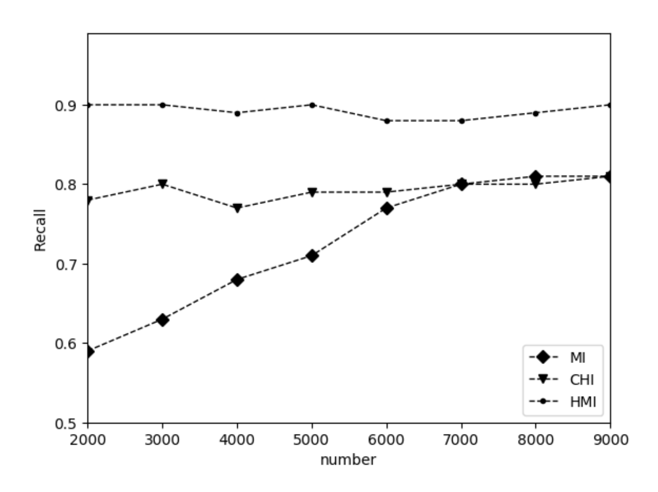


图 3　不同特征选择维度下召回率折线图

Fig. 3　Broken line diagram of recall rate under different feature selection dimensions

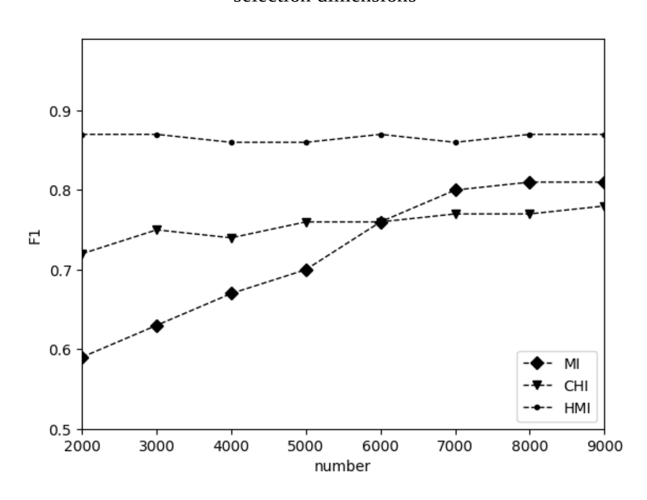


图 4　不同特征选择维度下 F1 值折线图

Fig. 4　F1-value line diagram under different feature selection dimensions

如图 5~7 所示，采用折线图来展现在美的空调评论数据集中不同维度下三个值的变化情况。

从图 5~7 可以看出，三种算法的准确率以及召回率都呈现上升的趋势，但明显 HMI 算法的整体效果要优于另外两种算法。另外，HMI 算法的 F_1 值在维度上具有较好的稳定性。

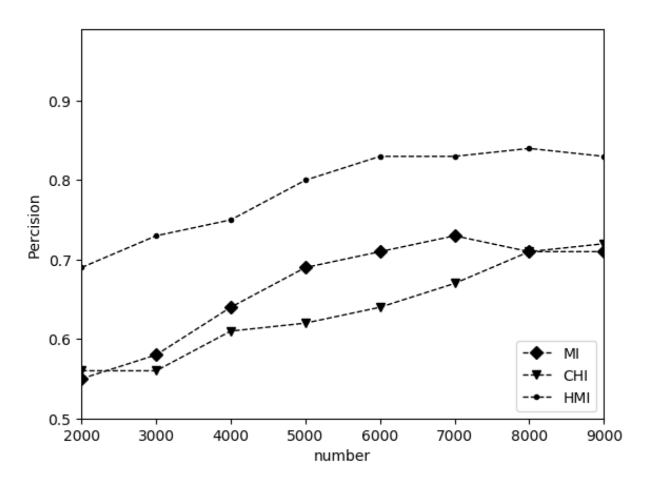


图 5　不同特征选择维度下准确率折线图

Fig. 5　Broken line diagram of accuracy under different feature selection dimensions

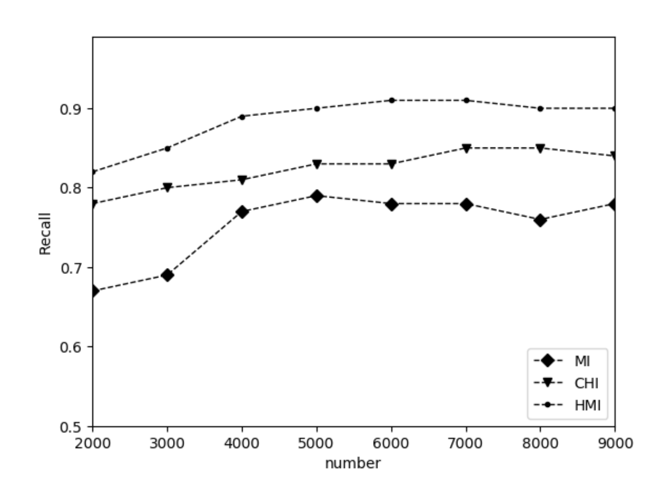


图 6　不同特征选择维度下召回率折线图

Fig. 6　Broken line diagram of recall rate under different feature selection dimensions

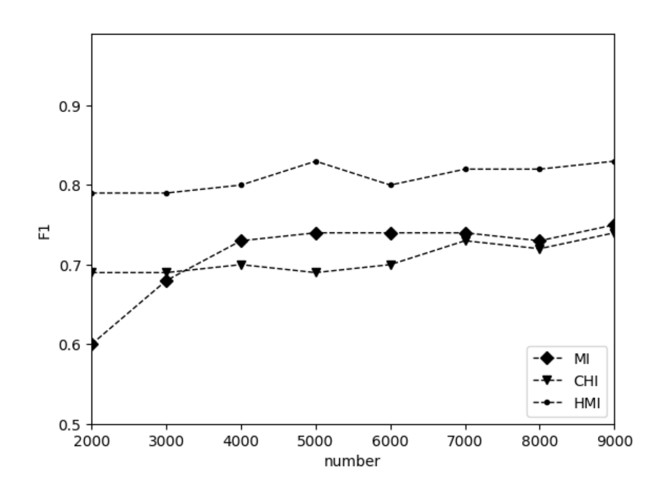


图 7　不同特征选择维度下 F1 值折线图

Fig. 7　F1-value line diagram under different feature selection dimensions

综上所述，混合互信息算法在准确率、召回率以及 F_1 值都明显高于 MI 和 CHI 方法，并且在特征维度的选择上具有较强的稳定性。因此可以得出，混合互信息（HMI）特征选择算法可以有效地提高特征选择的质量，进而提高文本情感分类的效果。

4　结束语

本文在分析互信息的特征选择方法存在的正负相关性现象以及忽略词频信息的问题基础之上，提出了一种混合互信息特征选择算法（HMI）。该算法通过引入逆文档频率，类间词频以及正负相关性指标，有效地使得词频信息在 MI 方法中得以有效地利用，并且很好地利用了正负相关性在该算法中不可忽视的作用。通过实验结果可以得出，混合互信息（HMI）方法明显优于其他特征选择方法，并在文本情感分类中取得了不错的效果。

参考文献：

[1] Cherry C, Mohammad S. Binary classifiers and latent sequence models for emotion detection in suicide notes [J]. Journal of Biomedical Informatics Insights, 2012, 5 (S1): 147-154.

[2] Atiyeh M, Hossein M M. Robust feature selection from microarray data Based on cooperative game theory and qualitative mutual information [J]. Advances in Bioinformatics, 2016, 2016 (1): 1-16.

[3] 李平，戴月明，王艳. 基于混合卡方统计量与逻辑回归的文本情感分析 [J]. 计算机工程, 2017, 43(12): 192-196. (Li Ping, Dai Yueming, Wang Yan. Text emotion analysis based on mixed chi-square statistics and logical regression [J]. Computer Engineering, 2017, 43(12): 192-196.)

[4] Tang Jian, Zhou Shuigeng. A new approach for feature selection from microarray data based on information [J]. IEEE/ACM Trans on Computational Biology and Bioinformatics. 2016. 13(6): 1004-1015.

[5] Bidi N, Elberichi Z. Feature selection for text classification using genetic algorithms [C]//Proc of the 8th International Conference on Modelling, Identification and Control. Piscataway,NJ:IEEE Press, 2016: 806-807.

[6] 朱颢东，陈宁，李红婵. 优化的互信息特征选择方法 [J]. 计算机工程与应用,2010,46(26):122-124. (Zhu Yidong, Chen Ning, Li Hongchen. Optimized mutual information feature selection method [J]. Computer Engineering and Application, 2010, 46 (26): 122-124.)

[7] 陶永才，赵国桦，石磊，等. 一种改进的 MapReduce 互信息文本特征选择机制 [J]. 小型微型计算机系统, 2018, 39(3): 433-438. (Tao Yongcai, Zhao Guohua, Shi Lei, *et al.* Improved MapReduce mutual information text feature selection mechanism [J]. Minicomputer System, 2018, 39(3): 433-438.)

[8] Li Kewen, Yu Mingxiao, Liu Lu, *et al.* Feature selection method based on weighted mutual information for imbalanced Data [J]. International Journal of Software Engineering and Knowledge Engineering, 2018, 28(8): 1177-1194.

[9] Coelho F, Braga A P, Verleysen M. A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems [J]. International Journal of Computational Intelligence Systems, 2016, 9(4): 726-733.

[10] 刘海峰，陈琦，张以皓. 一种基于互信息的改进文本特征选择[J]. 计算机工程与应用, 2012, 48(25): 1-4. (Liu Haifeng, Chen Qi, Zhang Yihao. An improved text feature selection based on mutual information [J]. Computer Engineering and Application, 2012, 48 (25): 1-4.)

[11] 林少波，杨丹，徐玲. 基于类别相关的新文本特征提取方法 [J]. 计算机应用研究, 2012, 29(5): 1680-1683. (Lin Shaobo, Yang Dan, Xu Ling. A new text feature extraction method based on category correlation [J]. Computer Application Research, 2012, 29(5): 1680-1683.)

[12] Calvo B, Larrariaga P, Lozano J A. Feature subset selection from positive and unlabeled example [J]. Patten Recognition Letters, 2009, 30(11): 1027-1036.

[13] Lin Yaojin, Hu Qinghua, Liu Jinghua, *et al.* Streaming feature selection for multilabel learning based on fuzzy mutual information [J]. IEEE Trans on Fuzzy Systems, 2017, 25(6): 1491-1507.

[14] Bostani H, Sheikhan M. Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems [J]. Soft Computing, 2017, 21(9): 2307-2324.

[15] 辛竹，周亚建. 文本分类中互信息特征选择方法的研究与算法改进 [J]. 计算机应用, 2013, 33(S2): 116-118. (Xin Zhu, Zhou Yajian. Research and improvement of mutual information feature selection method in text classification [J]. Computer applications, 2013, 33 (S2): 116-118.)

[16]